DECISION MAKING IN RESEARCH AND CLINICAL PRACTICE

Part II: The decision and its interpretation*

John Kramer and James Vargo
University of Alberta, Canada

This is Part II of a two-part paper. Part I laid the foundation for what is discussed here—decision making and its interpretation. Included are discussions of decision rules and their relevance to the research process, the meaning of decision errors, how to interpret tests of significance and how to distinguish between statistical significance and practical significance. An understanding of the concepts and techniques presented in Parts I and II of this paper will enable clinical therapists to evaluate more judicously the clinical importance of research papers in physical therapy as well as to better prepare them to design and conduct their own clinical research projects.

As a result of research in the medical and allied professions, it is incumbent upon practioners to continually keep pace with the advances in their respective professions. To do so requires an ability to critically digest, evaluate, and synthesize the research literature as it relates to their particular clinical interest. For physical therapists to maintain high-quality patient care, they must have an understanding of the rationale underlying the basic statistical techniques that are used to assess the effectiveness of various treatment procedures. Part I of this paper described some of these techniques, with particular reference to their meaning for the clinical situation. Part I laid the foundation for this part of the paper, which describes statistical decision making and its interpretation, the interpretation of tests of significance, and, of particular importance to the clinical therapist, the distinction between statistical significance and practical (or clinical) significance.

Decision rules

The level of significance specifies how improbable the observed test statistic must be before the null hypothesis is rejected (Hinkle *et al* 1979). The statements that designate the statistical conditions necessary for rejecting the null hypothesis are called decison rules. As such, decision rules are simply a formalization of the decision-making process based upon the level of significance (McCall 1975).

If the observed, or calculated test statistic has a low probability of occurrence, for example fewer than 5 times out of 100 (designated as $P<0.05$), then the researcher can reject the null hypothesis as probably not true. In this case, the calculated test statistic falls within the critical region under the curve. The interpretation is that the probability of obtaining a calculated test statistic of the magnitude observed is considered too small to be attributable to chance alone.

If the observed, or calculated test statistic has a high probability of occurrence, for example greater than 5 times out of 100 (designated as $P>0.05$), then the researcher fails to reject the null hypothesis. In this case, the calculated test statistic falls outside the critical rejection region of the distribution for the test statistic. The interpretation is that the probability of obtaining a calculated test statistic of the magnitude observed is sufficiently great that its occurrence can be attributed to chance alone.

For example, a physiotherapist wishes to compare the effectiveness of ice packs and hot packs with respect to increasing the range of motion of the knee joint. Patients are randomly assigned to each of the two treatments and the mean gain in range of motion (post-treatment minus pre-treatment range of motion) is determined for each group. If the obtained value of the calculated test statistic, which reflects the difference between the two group means, is greater than the critical value of the test statistic (at the selected level of significance for the utilized sample size) then the physical therapist would reject the null hypothesis of no difference between treatments. The conclusion would be that

John Kramer is currently Assistant Professor in Physical Therapy at the University of Alberta, Canada. He graduated in 1970 with a Bachelor of Science Degree. In 1974 he obtained his Masters in Physical Therapy and in 1979 he was awarded his PhD. His special professional interests include electrical stimulation as a strength improvement technique in normal muscle tissue, influence of the therapeutic modalities on sensory and motor nerve conduction and biomechanics in physical therapy He has written a number of articles for medical journals including: Physiotherapy Canada. The Canadian Journal of Applied Sciences *and* The Canadian Journal for Health Physical Education and Recreation.

James Vargo obtained his Bachelor of Arts (Honours) in 1968, Masters of Education in 1970 and his PhD in 1972 at the University of Alberta, Canada He is an Associate Professor at the Department of Occupational Therapy at the University of Alberta. His special professional interests include counselling the disabled, psychological aspects of illness, and disability and changing public attitudes toward the disabled. He has written a number of articles for both physical therapy and occupational therapy journals in America, Canada and New Zealand.

* *Part I: 'The basis for decision' was published in the June issue of this journal.*

the two treatments were probably different in their effects on the range of motion of the knee joint. If, on the other hand, the obtained value of the calculated test statistic was less than the critical value, the physical therapist would not reject the null hypothesis. This decision infers that there is not sufficient evidence to say the two treatments probably differ in their ability to improve range of motion of the knee joint.

Decison errors

Hypothesis testing is based on probability theory. There always exists a possibility of having made the wrong decision. In any statistical test of significance, two correct decisions are possible and two incorrect decisions (decision errors) are possible. Type 1 error, or alpha error, refers to a decision to reject the null hypothesis, when it is in fact true. Alpha also refers to the level of significance of the test of significance and, consequently, can be thought of as the probability of rejecting the null hypothesis by mistake. That is, a decision is made that there is a difference when no difference actually exists. Type 2 error, or beta error, refers to a decision not to reject the null hypothesis, when it is in fact false. That is, the test of significance fails to detect a true difference between treatments. Figure 1 illustrates the relationship between the true situation and the decision made (Ferguson 1971, Hinkle *et al* 1979, Huntsberger and Leaverton 1970, Siegel 1956).

If researchers decrease the level of significance, for example from 0.05 to 0.01, they are being more conservative. Although there is now less likelihood of incorrectly rejecting a true null hypothesis, small differences could go undetected. On the other hand, if researchers increase the level of significance, for example from 0.01 to 0.05, they are being more liberal. Although there is now a greater likelihood of detecting small differences, there is an increased probability that some of these differences are attributable to chance occurrence and are not real, or true, differences (Colton 1974, Hinkle *et al* 1979, McCall 1975, Currier 1979). The investigator selects the level of significance with these decision errors in mind.

Interpretation of tests of significance

Tests of significance concern themselves with the likelihood of occurrence of an observed sample value, if the null hypothesis represents the true situation. Simply declaring a difference between treatments to be statistically significant or not statistically significant is inadequate. Statistical significance infers that an observed difference between treatments probably did not arise by chance and is a true (or a real) difference. In other words, the outcome of statistical significance indicates that the observed sample value is extremely improbable if the null hypothesis is true. Lack of statistical significance indicates that an observed difference between treatments is probably attributable to chance occurrence and is not a true (or a real) difference (Huntsberger and Leaverton 1970).

The decision as to how likely, or probable, an observed event is, is based on the level of significance. This value is arbitrarily selected by the researcher and has considerable bearing on the statistical decision. After the statistical evaluation, one still does not have proof of a correct decision, but only a probability that a correct decision, relative to the the null hypothesis, was made. A decision to reject the null hypothesis does not mean that the alternative hypothesis has received an equivalent degree of support (Neale and Liebert 1973). For example, if the null hypothesis were rejected at the 0.05 level of significance ($P<0.05$), this fact cannot be interpreted to mean that the researcher is 95 per cent certain that the alternative hypothesis is true. Rather, the appropriate interpretation is that the observed result is highly unlikely if the outcome were the product of chance factors alone. Statistically, the investigator is justified in suggesting that the alternative hypothesis is a more likely explanation for the observed result.

Aside from interpreting the numerical evidence related to the occurrence of a particular event, other factors must be considered by both the researcher and the reader. There may be rival explanations, other than the alternative hypothesis, that are viable explanations for the observed result. These factors may, or may not, be recognized by the researcher, for example age, sex, intelligence, occupation, and could offer even more likely explanations for the observed result. It is the responsibility of the consumer of the research to decide to what extent such factors may have influenced the observed result. Only then can the consumer decide what level of confidence to place in the investigation.

For example, suppose that the physical therapist compares the range of motion of trunk flexion of patients before and after a low-back exercise programme. An appropriate test of significance indicated that the difference between pre-exercise programme scores was statistically significant. The physical therapist interprets this to mean that the patients improved significantly as a result of the exercise programme. However, a rival explanation could be offered; namely, that the improvement occurred as a result of the passage of time, and, therefore, patient mobility would have improved even without the exercise programme. One method of discounting this rival explanation would be to compare the scores, before and after a corresponding time period, for a control group—similar patients who did not receive the particular exercise programme. In this instance, a statistically significant difference between the post-exercise

scores for the two groups would warrant the conclusion that the exercise programme did in fact improve the range of motion of trunk flexion.

Statistical significance and practical significance

Statistical significance refers to the relationship of a calculated test statistic (or a sample statistic), based on sample data, with a critical value for the test statistic (or a population value). This critical value is based on a theoretical population model and corresponds to the selected level of significance. If the likelihood of obtaining a claculated test statistic as large as that obtained is equal to or less than the level of significance, then the test statistic is termed to be statistically significant. Statistical significance relates to the likelihood of obtaining a test of a particular magnitude as a result of chance factors.

The decision that the results of a test of hypothesis are statistically significant is not synonymous with the conclusion that differences of great practical importance exist (McCall 1975). Once statistical significance has been found, the investigator must still decide if this statistical difference really makes a practical difference. The answer to the question of practical importance cannot be statistically determined, but depends on the research consumers' knowledge of the significance of implementing or failing to implement the information. Exactly what is important practically is difficult to define because it is not a point of fact, but depends on one's point-of-view and the current situation.

An observed difference between treatment effects may be statistically significant, but not of practical importance. Conversely, a real and important practical difference may exist between treatment effects despite the fact that the results do not achieve statistical significance (Huntsberger and Leaverton 1970). For example, a standard post-knee-menisectomy programme may involve isometric exercises, while an alternative programme may include isometric exercises in combination with electrical stimulation. A clinician may wish to know if the extra time and expense associated with the administration of the particular electrical stimulation programme is justified when compared with that of the standard programme. To answer this question, the therapist might standardize the isometric exercise and the electrical stimulation programme at several hospitals within a large city. Assume that, after one year, 500 patients have participated in each programme and the strength scores (straight leg raise weight lifted following one week hospitalization) were compared by an appropriate statistical test of significant ($P<0.01$), and a 0.25 kg difference in favour of the electrical stimulation group was found. Clinically, this is of little practical value. Although there is statistical significance, this 0.25 kg difference does not appear to warrant the extra time and expense associated with the electrical stimulation treatment. Despite the statistical significance, the therapist must still ask 'of what value is the extra 0.25 kg?' Statistics cannot answer this question. The statistical test simply suggests that this difference probably did not come about as a result of chance factors.

To answer the same question, another therapist, who works in a small clinic, develops a similar study. However, practical considerations of time and expense restrict the sample to patients referred to the clinic for treatment. Assume that, after 6 months, five patients have participated in each programme and that the strength scores are compared by an appropriate statistical test of significance. Assume, in this case, that the difference between the means of the two groups was 3 kg and that this difference was not statistically significant. Clinically, the 3 kg difference could be of practical importance. The electrical stimulation group, having achieved the highest strength level may, as a result, achieve normal knee function and normal gait more rapidly. However, in this study the small sample size (five patients per group) made it more difficult to obtain a statistically significant test statistic. The physical therapist is still faced with a decision—is the 3 kg difference of any practical importance, whether or not it is statistically significant at a particular level of significance? It should be emphasized that the probability of a given event occurring by chance is dependent on sample size (Hinkle *et al* 1979, Huntsberger and Leaverton 1970, Neale and Liebert 1973). In the above example, a larger sample size might have permitted the attainment of statistical significance.

The point to be emphasized here is that statistical significance relates to the likelihood of the occurrence of an observed result, not to its practical importance. The fact that a test statistic is statistically significant indicates nothing about its practical importance or the many possible contributory variables that could be responsible for the observed outcome, such as the format of the electrical stimulation current, sex or age of the subjects, the surgeon, the therapist.

The previous discussion illustrated one case in which there was statistical significance, but little, or no, practical importance, while the second case illustrated practical significance, despite the lack of statistical significance. Researchers sometimes fail to recognize this apparent contradiction. The failure to find a statistically significant difference between treatments is dreaded by many researchers. However, the observation of a non-statistically significant difference between treatments can be of considerable practical importance. For example, in the comparision of several treatment programmes, the observation of no statistically significant difference between treatments can be clinically useful by suggesting that the treatments

are equally effective. When such is the case, the therapist can select treatment based on other factors such as time, expense and equipment. The attainment of statistical significance lends support to the position that the observed result was not attributable to chance factors, but to a true difference. For this reason, the attainment of statistical significance is usually desirable and lends confidence to the researcher's postion.

Conclusions

It should be emphasized that statistical procedures do not improve the quality of information. They do not correct for inadequate or incorrect data-gathering procedures and they do not guarantee that a correct decision will ultimately be made. Numbers by themselves do not constitute a full analysis of information, and all statistical procedures still require a verbal interpretation.

Statistical significance and practical significance are not dependent on one another. Statistical significance implies that a difference between treatments, as observed under the specific conditions of the investigation, most likely did not occur by chance. Practical significance, on the other hand, relates to the application of knowledge gained through research. If one can have confidence that a particular observed difference between treatments probably did not arise as a result of chance factors, then that difference can justifiably be considered as real and thereby incorporated into future clinical plans A difference between treatments that is not statistically significant provides little confidence that it is a true difference. Such an observed difference is probably the result of chance occurrence, rather than a true difference.

Statistical techniques objectify the decision-making process by establishing decision rules. These techniques do not free the researchers or the clinician from making the ultimate decision on the practical use of such information.

References

Colton T (1974), *Statistics in Medicine,* Little, Brown and Company, Boston.

Currier D P (1979), *Elements of Research in Physical Therapy,* Williams and Wilkins, Baltimore.

Ferguson G A (1971), *Statistical Analysis in Psychology and Education,* McGraw-Hill, New York.

Hinkle D E, Wiersma W and Jurs S G (1979), *Applied Statistics for the Behavioral Sciences,* Rand McNally, Chicago.

Huntsburger D V and Leaverton P E (1970), *Statistical Inference in the Biomedical Sciences,* Allyn and Bacon, Boston.

McCall R B (1975), *Fundamental Statistics for Psychology* (2nd edn) Harcourt, Brace, Jovanovich, New York.

Neale J M and Liebert T M (1973), *Science and Behavior: An Introduction to Methods of Research,* Prentice-Hall, Inc, Englewood Cliffs, New Jersey.

Siegel S (1956), *Non-parametric Statistics for the Behavioral Sciences,* McGraw-Hall, Toronto.

Figure 1.: The relationship between the actual situation, and the decision made and the possible decision errors associated with all tests of significance